



Les conditions de la connaissance de soi

Joëlle Proust

► To cite this version:

Joëlle Proust. Les conditions de la connaissance de soi. Philosophiques, 2000, 27 (1), pp.161-186.
ijn_00139335

HAL Id: ijn_00139335

https://hal.science/ijn_00139335

Submitted on 30 Mar 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Les conditions de la connaissance de soi
in *Philosophiques* 27,1,2000, 161-186.

Joëlle Proust

CNRS

CREA, Ecole Polytechnique

Résumé

La connaissance de soi suppose que l'on puisse former des pensées vraies de la forme "je Y que P", où "Y " fait référence à une attitude propositionnelle, "P" à son contenu, et "je" au penseur de cette pensée. La question qui se pose est de savoir, ce qui, dans le contenu mental occurrent [P], justifie l'auto-attribution de cette pensée. Ce problème dit de la transition soulève trois difficultés ; celle de la préservation du contenu intentionnel entre la pensée de premier et de second ordre ; celle de la reconnaissance de l'attitude ayant ce contenu intentionnel pour objet, et enfin la reconnaissance que ce qui est pensé l'est par le sujet qui pense. Le présent article se propose de montrer que la troisième difficulté résiste à une approche fondée sur l'expérience ou sur la signification cognitive de [P], et avance l'idée que la notion d'action mentale permet d'éclairer les conditions d'identité du penseur de [P] et du sujet de l'auto-attribution de l'attitude propositionnelle "Y que P".

Abstract

Self-knowledge requires the capacity to think true thoughts of the form "I Y that P", where "Y " refers to a propositional attitude, "P" to a propositional content, and "I" to the thinker of the thought of content P. The question that this requirement raises is to know what, in the occurrent mental content [P] , justifies the self-attribution of this thought. This problem, called the transition problem, includes three difficulties : how is intentional content stable across first- and second-order thoughts ? How is the attitude with this intentional content identified by the thinker ? And how is the thinker of the second-order thought able to claim truly that he himself is the thinker of the first-order thought ? The present paper's aim is to show that the third difficulty cannot be solved through an examination of the experience of having [P] or on the basis of the cognitive significance of P, and suggests that an analysis of mental actions in which propositional attitudes play a causal and feedback role give a better grasp on the conditions of identity of the thinker of [P] with the thinker of the propositional attitude "Y that P".

Les conditions de la connaissance de soi

Joëlle Proust

CNRS

CREA, Ecole Polytechnique

L'un des problèmes philosophiques importants que pose la question des relations entre la première personne et l'auto-attribution d'états mentaux concerne le rapport entre la métaphysique et l'épistémologie

de la première personne. Toute forme de réductionnisme psychophysique doit s'atteler à la tâche de savoir non seulement sur quelles connaissances un sujet s'appuie pour savoir qu'il est un sujet, ou de quelles sources il tire cette conviction, et déterminer la nature du processus épistémique qui produit cette connaissance ou cette conviction. Il doit également se prononcer sur la nature ontologique du sujet; s'agit-il d'une *propriété* mentale, (comme celle d'avoir conscience d'un état intentionnel, d'avoir des impressions qualitatives ou d'avoir co-conscience de plusieurs états), ou d'un *individu* capable de conscience, propriétés et individus bien déterminés pour lesquels il convient de rechercher de quelles structures physiques ils dépendent systématiquement ?

Quelle que soit la réponse apportée, on peut supposer que le philosophe naturaliste sera peu tenté de considérer l'articulation de ces deux questions - la question épistémologique de la manière dont on acquiert les manières de penser à soi comme sujet, et la question ontologique de ce qui rend vraies les pensées sur soi - comme un problème d'intégration, selon le terme de Christopher Peacocke. L'intégration suppose l'existence d'intuitions indépendantes qu'il s'agirait de réconcilier, comme par exemple c'est le cas pour le problème mathématique (ou théologique) de l'infini. Comment l'esprit humain, ou tout autre esprit fini, peut-il concevoir l'infini ? L'intuition que partagent beaucoup de philosophes naturalistes à propos du sujet conscient, c'est que la métaphysique du sujet n'est pas séparable, comme le sont d'autres objets de connaissance, de la manière dont le sujet est appréhendé, que ce soit par lui-même ou par les autres sujets. Cette intuition concernant le type de lien qui unit le sujet et le processus de sa construction mentale demande évidemment à être affinée : s'agit-il d'un processus épistémique pleinement justifié, ou seulement fiable, permettant de faire référence à un sujet réel qui serait le porteur de ses propriétés mentales ? Ou bien s'agit-il d'un processus illusoire, qui se développerait à la faveur d'une sélection étrangère à un critère de vérité, sans que le sujet puisse jamais former l'objet d'une référence fondée, ni se voir attribuer de manière vraie des propriétés qui lui appartiendraient objectivement ?

Selon le type de jugement que l'on porte sur la genèse épistémique, diverses positions métaphysiques sur la nature des sujets pourront être distinguées. Celle du *réalisme* du sujet, - entité stable ou dynamique - conçu au terme d'une genèse cognitive. Celle du *réductionniste*, qui dissout le sujet au profit d'une propriété particulière des états mentaux sous-jacents. Une fois seulement ces thèses métaphysiques établies, sera-t-il possible de rechercher la base de survenance des entités ou propriétés subjectives reconnues par la théorie. L'objectif du présent article est plus modestement d'établir les conditions (métaphysiques et/ou épistémiques), si elles existent, auxquelles un sujet conscient peut s'attribuer ses propres contenus mentaux d'une manière qui soit stable dans le temps, et qui résiste à la diversité des contenus mentaux et des attitudes qui les régissent.

1 - L'usage subjectif du *je*

Pour savoir si un sujet peut atteindre la connaissance de soi, il convient d'analyser le type de situation dans laquelle la référence au sujet intervient pour ainsi dire "à l'état naissant", dans des circonstances où une expérience subjective, c'est-à-dire consciente et pourvue de propriétés qualitatives, suscite directement un jugement auto-attributif. Cet usage du "je" se distingue de l'usage référentiel ordinaire des termes singuliers dans le langage public, où le terme "Je" fonctionne comme un terme singulier qui renvoie à l'auteur de l'énoncé. Une occurrence d'énoncé dont le sujet est "je" implique nécessairement l'*existence* de l'objet auquel "je" fait référence, et permet également aux interlocuteurs d'*identifier* l'objet auquel "je" fait référence, étant donné le contexte de la communication. Dans la pensée en revanche, et en dépit de l'argument cartésien à cet effet, il n'est pas évident que de la simple perception occurrente d'un état de chose ou de la pensée que *P*, puisse être dérivée l'existence du sujet qui pense cette pensée, et encore moins que l'existence d'un tel sujet s'ensuive nécessairement. On peut en effet opposer à la déduction cartésienne que du simple fait qu'un contenu mental soit appréhendé (cru, désiré, ressenti, etc.), on peut certes dériver qu'un événement mental se soit produit (un physicaliste ajoutera qu'un événement cérébral identique à l'événement mental se soit produit). Mais la proposition qui permet d'analyser cet événement pourrait être de la forme impersonnelle "il pense", par analogie avec "il pleut", comme Lichtenberg, repris par Wittgenstein, l'a soutenu. En d'autres termes, de la simple existence d'une pensée occurrente, il ne semble pas légitime de déduire qu'un sujet qui pense existe nécessairement, pas

plus que de l'existence d'un événement physique, on ne peut dériver l'existence nécessaire d'un agent qui en serait la source intentionnelle.

Cet exemple nous donne l'occasion de prévenir un parallèle trop rapide sur lequel nous aurons l'occasion de revenir dans la suite. "Il pleut" se dit dans une situation sans agent, où l'information et l'intentionnalité ne jouent aucun rôle. Le cas de la relation entre la pensée et son éventuel porteur n'a pas à être antisymétrique à celui de la pluie. Si l'on suppose que toute pensée occurrente implique nécessairement un sujet, ce n'est pas au sens où le penseur serait *l'agent* de sa pensée occurrente. La relation entre le penseur et sa pensée doit être pour le moment laissée dans un certain flou; la littérature caractérise cette relation par le terme de "possession".

Les philosophes qui ont réfléchi sur l'auto-attribution des états mentaux ont proposé de caractériser la situation "naissante" de l'usage en pensée du *je* de manière à isoler les cas où le sujet s'attribue une pensée qu'il est le seul à pouvoir s'attribuer. Il le fait, dans ce cas, avec une autorité qui n'appartient qu'à lui. Wittgenstein a ainsi opposé les usages du *je comme objet* des usages du *je comme sujet*. Dans l'usage objectif, l'usage référentiel du mot "je" est fait en vue de permettre l'identification du référent par son aspect physique, c'est-à-dire sur une base essentiellement publique. Des exemples en sont fournis par "J'ai grandi de 12 centimètres", "j'ai une bosse sur le front", etc. Dans l'usage subjectif, comme dans "j'ai mal aux dents", ou "j'essaie de lever le bras", "je" n'a pas pour fonction première d'identifier une personne particulière. Ce qui l'indique, c'est qu'il est impossible de se tromper de personne quand on dit "j'ai mal aux dents". Alors que l'usage objectif implique la possibilité d'une fausse reconnaissance, (je croyais me voir dans le miroir, mais c'est mon jumeau que je vois), la fausse reconnaissance est absurde dans le cas de l'usage subjectif ; la possibilité de l'erreur sur la personne est *a priori* exclue. Un sujet qui s'attribue un état mental bénéficie, en d'autres termes, d'une complète "immunité à l'erreur d'identification". Expliquer la connaissance du sujet acquise par auto-attribution d'états et d'événements mentaux suppose ainsi que l'on statue sur cette immunité.

Pour Wittgenstein, cette immunité renvoie à un fait de langage : le "je" de "je souffre" n'est pas selon lui un démonstratif ; il ne fait pas référence à la personne qui parle ou qui ressent au sens où cet usage du *je* aurait une valeur contrastive par rapport à d'autres référents possibles :

Comment savez-vous que vous souffrez ? -- "Parce que je le *sens*." Mais "je sens" a exactement le même sens que "je souffre". Il ne s'agit donc nullement d'une explication. Le fait cependant que, dans ma réponse, j'appuie sur le mot "sens" et non pas sur le "je" indique bien que par ce "je" je n'entends pas distinguer une certaine personne parmi d'autres.

Les auto-attributions subjectives sont pour Wittgenstein de fausses affirmations sur soi. Elles portent non sur la personne qui dit "je", mais sur l'état ressenti. Ne portant pas d'information véritable venant s'ajouter au contenu d'expérience correspondant, elles sont comparables à des tautologies comme "ici est ici". Rien de substantiel n'est dénoté par le terme singulier "je". Ce qui fait la force de l'argument de Wittgenstein, c'est que toute interrogation sur l'identité paraît généralement présupposer la référence à un "je" immunisée contre l'erreur : en posant la question "est-ce moi qui y ?", le penseur cherche non pas à s'identifier lui-même, mais à établir l'identité entre lui-même et la personne qui instancie la propriété mentale y.

Quelles conclusions peut-on tirer toutefois de cette remarque ? Le cas de l'auto-attribution ne diffère pas ici du cas général de la catégorisation d'objet : pour s'interroger sur l'identité d'un objet, il faut non seulement disposer d'un ensemble de concepts permettant de le catégoriser, mais aussi avoir la capacité de distinguer l'objet auquel on fait référence. De même que l'identification d'un objet suppose qu'on le discrimine avant de lui attribuer une propriété, le "je" ne peut devenir le thème d'auto-attributions que s'il est déjà distingué parmi d'autres. Appelons identification primaire la discrimination qui donne l'accès à la référence à soi, et identification secondaire, ou réidentification, les usages ultérieurs du "je".

La manière de résister à la thèse de Wittgenstein sur le caractère tautologique de la référence à "je" consiste à maintenir que l'identification secondaire paraît automatiquement immunisée contre l'erreur parce que l'identification primaire est supposée acquise. L'agent qui, pour des raisons développementales ou psychopathologiques, échoue dans l'identification primaire, n'a pas de connaissance de soi en mode subjectif parce que, même s'il jouit d'expériences subjectives, il est incapable de faire référence à soi. C'est donc peut-être simplement parce qu'il se concentre sur les usages corrects - secondaires - des pensées subjectives en "je" que Wittgenstein constate que ce n'est pas sur le "je" que porte généralement le contenu de connaissance pertinent.

Il existe en outre des circonstances où l'on peut légitimement se poser le problème de l'identification secondaire : se demander si l'on est *soi-même* le porteur d'un état mental ou d'une attitude propositionnelle. Il est à cet égard frappant de remarquer que l'immunité à l'erreur d'identification ne s'applique pas aux souvenirs en première personne. Comment peut-on adéquatement délibérer, s'engager dans des processus de révision de croyance, résoudre des problèmes, sans avoir la possibilité de répondre de manière véridique à des questions comme :

est-ce moi qui ai eu cette idée ?

est-ce moi qui ai rêvé que *P*, ou est-ce toi qui me l'as raconté ?

Dans ce type de rapport subjectif, il devient possible pour le sujet de se méprendre sur ce qui a été, à une époque antérieure, son propre contenu mental. Si la réponse à ces questions *au passé* est source de connaissance, il semble plausible de dire qu'il doit exister une proposition correspondante *au présent* qui est également source de connaissance pour le sujet qui la pense.

Si l'on souhaite maintenir contre Wittgenstein que la référence à soi-même vise bien à distinguer une personne, il semble qu'on doive soit rejeter le principe d'immunité à l'erreur d'identification : dans ce cas, la référence à soi cesse d'être tautologique, et peut donner lieu à une connaissance substantielle ; soit montrer que l'immunité à l'erreur d'identification ne vaut que dans certaines conditions, et non de manière universelle; elle ne découle pas d'un principe *a priori*, mais du fonctionnement normal d'un organisme capable de former des pensées en "je". Les deux solutions ont en commun de nier que l'immunité à l'erreur d'identification dans l'usage subjectif de "je" procède d'un principe *a priori*. La présente stratégie consiste à adopter la seconde solution, en défendant l'idée que l'identification primaire détermine bien un champ de connaissance sur soi. L'immunité à l'erreur n'est pas nécessairement le signe d'une vacuité référentielle du constituant *Je* dans ses usages non publics. Examinons maintenant de plus près la transition qui s'opère dans l'auto-attribution des contenus mentaux par un sujet.

2 - Le problème de la transition

La transition entre un contenu mental et son attribution fait problème dans le cas où la référence à soi dans l'auto-attribution est de type "subjectif". Dans ce cas, nous n'avons pas de moyen d'identification physique, par exemple le repère du corps individuel, ni de point d'appui sur une référence partagée (comme "celui qui vous parle"). Le contenu mental occurrent de la pensée ne fournit apparemment pas en lui-même d'information sur laquelle l'auto-attribution pourrait se fonder. Prenons par exemple la pensée perceptive occurrente :

(1) [Le téléphone est sur la table]

Comment le penseur peut-il, à partir de cette occurrence, former la pensée auto-attributive :

(2) [Je vois que le téléphone est sur la table] ?

Répondre à cette question se heurte d'emblée à trois difficultés liées au fait que nous n'avons pas encore

suffisamment précisé le cadre théorique dans lequel nous la posons. La première difficulté consiste à savoir de quoi le contenu intentionnel de (1) est constitué. S'agit-il d'un contenu portant sur des *objets*, sur une *propriété* particulière -- leur relation spatiale -- ou sur le *fait* que ces objets ont cette propriété ? La seconde difficulté consiste à préciser la nature du contenu de la perception : s'agit-il du contenu non-conceptuel qui sert de base perceptive au jugement perceptif correspondant ? S'agit-il d'un contenu conceptuel articulant une propriété (ou une relation) et un (ou deux) individu(s) ? La troisième difficulté est d'établir de quoi nous sommes conscients quand nous avons un contenu mental comme (1) ; est-ce que les seules caractéristiques dont nous sommes conscients sont intentionnelles ou représentationnelles, ou bien l'expérience a-t-elle des traits qualitatifs non représentationnels qui détermine l'"impression que cela fait" d'avoir cette expérience ?

Nous ne pouvons dans le cadre du présent article examiner toutes les possibilités que les divers types de réponses à ces trois questions déterminent. Nous déciderons sans pouvoir justifier ce choix ici que le contenu mental de (1) consiste dans un fait, c'est-à-dire un état de chose réalisé, impliquant en l'occurrence une certaine relation spatiale, [sur], entre deux individus, [le téléphone] et [la table]. Le fait qui forme le contenu mental en question est exprimé, nous le supposons également, dans un format représentationnel conceptuel, ancré dans une présentation visuelle qui en fournit le contenu non conceptuel. En d'autres termes, le contenu mental consiste dans le jugement perceptif qui fait suite à l'extraction d'une information visuelle portant sur les relations spatiales entre la table et le téléphone.

La troisième question, en revanche, ne peut être réglée aussi rapidement. Car savoir *de quoi* nous sommes conscients quand nous avons une pensée dont le contenu est (1) est crucial pour déterminer si (1) contient déjà les éléments d'une transition vers (2). Deux possibilités extrêmes se présentent. La première possibilité consiste à explorer la nature phénoménologique du contenu de conscience. Les deux individus constituant de (1) ont des propriétés telles que la couleur ou la forme, et pour l'un des deux, occasionnellement, le timbre, la fréquence, le volume du son, autant de propriétés qui sont tenues par cette théorie essentiellement inhérentes à la conscience que j'en ai, et seulement indirectement liées à l'objet intentionnel. Dans cette perspective, je ne peux pas former le jugement perceptif que le téléphone est sur la table sans avoir recouru à la phénoménologie de mon expérience, c'est-à-dire aux propriétés de mon état de conscience lorsque je perçois le téléphone sur la table.

Dans ce cas, la transition entre le contenu intentionnel (1) et l'auto-attribution de (1) dans (2) comporte trois étapes : i) j'ai une expérience phénoménologique ; ii) je découvre à partir d'elle (par inférence) le contenu intentionnel de mon expérience ; iii) j'utilise les moyens conceptuels dont je dispose pour décrire mon rapport épistémique à ce contenu. En particulier, pour connaître le type d'expérience que j'ai dans (i), je dois savoir que la couleur ou la forme me sont données visuellement.

L'autre possibilité extrême serait que seul le contenu (1) lui-même, c'est-à-dire une relation spatiale particulière entre deux individus, forme le contenu intentionnel conscient de la pensée correspondante. Ce dont le sujet est conscient est dans ce cas non *l'expérience* de (1) ou ses propriétés, mais le contenu intentionnel de (1). Cette théorie, selon laquelle la conscience phénoménale est transparente, et n'a pour contenu rien d'autre que ce qui est intentionnellement représenté par l'état mental correspondant, est défendue par les tenants de la nature strictement représentationnelle des *qualia*. Pour savoir l'effet que cela fait d'avoir l'expérience de (1), il suffit de regarder ce qui forme l'objet de cette expérience.

Dans ce cas, la transition entre le contenu intentionnel (1) et l'auto-attribution de (1) par (2) est indirecte et médiate ; c'est non pas en me fondant sur mon expérience de (1) que je peux connaître (2) puisque (1) porte sur un état de chose extérieur, et non sur un état interne. Dans la perception du fait que le téléphone est sur la table, ce que je perçois directement, c'est l'état de chose extérieur, et non pas l'expérience que j'en ai. Mais comme les propriétés des objets extérieurs sont récurrentes, et que je sais "à quoi ressemble le téléphone", la connaissance que j'ai de mon expérience du téléphone a elle aussi toutes les apparences de l'immédiateté. Elle est pourtant *inférée*. Voyons de quelle manière. Je typifie mes expériences sur la base de leur contenu intentionnel. Mon expérience visuelle présente est celle que j'ai chaque fois que je vois le téléphone. Pour pouvoir ainsi typifier les *contenus d'expérience*, il faut avoir les concepts permettant de subsumer les individus, propriétés et relations présents dans les contenus d'expérience ;

pour typifier les *attitudes propositionnelles*, il faut posséder les concepts subsumant le type d'opération mentale (ou d'attitude) qui prend pour objet le contenu considéré. Ainsi, reconnaître et réidentifier un téléphone suppose que l'on ait le concept de téléphone. Reconnaître la présence d'une expérience visuelle quand je perçois le téléphone sur la table suppose que je dispose du concept de vision.

En ce qui concerne les expériences internes, telles que "ressentir un piquêre sur le doigt", la même explication peut être utilisée. L'expérience que j'ai porte sur l'état physique de mon doigt, dont la peau a subi l'intrusion d'un objet pointu. C'est ce fait qui est consciemment perçu et non à proprement parler la douleur ; dire que "j'éprouve une douleur" dépend conceptuellement de l'état intentionnel que j'ai en percevant un corps pointu au contact de mon doigt. Je peux éprouver une douleur sans avoir le concept de douleur. Mais je ne peux pas m'auto-attribuer une expérience occurrente de douleur sans avoir, d'une part, le concept de doigt, de piquêre, et d'autre part, celui de douleur.

3 -La transition comme lien nécessaire

Même s'il y a en fait de bonnes raisons de préférer la conception représentationnaliste des propriétés phénoménales (celles qui constituent la conscience que le sujet percevant a de son environnement), il n'est pas nécessaire de trancher entre les deux versions opposées du contenu mental pour notre présent objectif, qui consiste à analyser la transition entre (1) et (2). Nous allons voir qu'en fait, dans l'une comme dans l'autre manière de présenter la question, le même problème se pose : la transition entre (1) et (2) n'est possible que si l'on effectue une pétition de principe sur la justification ontologique de cette transition, c'est-à-dire si l'on présuppose l'identité entre le penseur qui se représente (1) et le penseur qui s'auto-attribue cette pensée dans (2).

Pour faire apparaître la difficulté, commençons par généraliser la formulation de notre question : étant donné un contenu intentionnel occurrent P donnant lieu à une attitude propositionnelle occurrente de type Y, comment le sujet peut-il former de manière autorisée l'auto-attribution correspondante :

(3) Maintenant [P] (de type Y)

(4) Maintenant je Y que P

L'hypothèse de la nécessité métaphysique

Une premier type de solution mérite d'être examiné. N'y a-t-il pas entre ces deux pensées un lien nécessaire, qui justifie que la transition vaille comme connaissance de soi ? Examinons en premier lieu si ce lien est d'ordre métaphysique, c'est-à-dire ancré dans la nature même des états mentaux concernés, et dans l'existence d'un rapport intrinsèque liant le premier au second ordre. Ecartons d'emblée un argument ici sans pertinence. Je peux avoir le contenu mental [P] sans avoir l'attitude de type Y ; il n'y a pas de nécessité métaphysique à ce que [P] soit vu maintenant, plutôt que simplement imaginé ou souhaité. Mais cet argument ne vaut pas dans le cas de (3) , où le *type* de P est fixé du simple fait que P est un état mental occurrent. Le problème du lien métaphysique entre (3) et (4) s'attache à l'existence de deux états mentaux, ou à l'existence d'un état mental et à l'existence d'un "je" auquel cet état mental conduirait nécessairement, comme dans la dérivation cartésienne qui part d'un état de conscience ayant un contenu déterminé (penser ou imaginer, ou percevoir que P) pour conclure à l'existence d'un je qui pense, imagine, ou perçoit que P.

L'une des façons de défendre l'existence d'un tel lien métaphysique entre (3) et (4) pourrait consister à soutenir qu'il existe une relation constitutive entre eux. On pourrait par exemple défendre l'idée que les états mentaux d'ordre inférieur tels que (3) activent nécessairement un état mental d'ordre supérieur tel que (4) parce qu'ils en sont constitutivement inséparables. Dans cette hypothèse, soutenue entre autres par Davidson, les états mentaux d'ordre inférieur sont ainsi constitués en partie par les états d'ordre

supérieur. On pourrait également défendre l'hypothèse converse, selon laquelle les états mentaux d'ordre supérieur sont en fait constitués en partie par les états d'ordre inférieur. Dans les deux cas de figure, la thèse métaphysique portant sur le rapport de constitution entre les états mentaux (3) et (4) fonde la thèse épistémologique selon laquelle appliquer un concept quelconque suppose que l'on ait la capacité de second ordre d'appliquer le concept de concept : le contenu *f* implique le contenu *Y* (*f*), ou réciproquement.

Avant de discuter la portée de cette thèse pour résoudre le problème de la transition, il convient de distinguer les *trois* manières dont cette dérivation peut être évaluée, et qui donnent lieu à trois types de questions. D'abord, est-ce que le contenu [P] qui forme l'objet intentionnel est correctement reconnu par le sujet comme étant l'objet de sa pensée? Dans les termes de Peacocke, existe-t-il une "sensibilité inter-niveaux" à la stabilité du contenu ? En second lieu, est-ce que les types d'opération ou d'attitude mentales qui sont effectivement exercés de manière occurrente à propos de [P] sont correctement identifiés par le sujet comme étant les opérations ou attitudes qui sont les siennes à propos de [P] ? Enfin et surtout, le sujet est-il fondé à considérer qu'il peut s'auto-attribuer l'état mental en question ? Le "je" qui intervient dans (4) est-il légitimé par les éléments de (3) ?

Résumons ces trois types de sensibilité à l'état mental qui doivent être conjugués pour qu'une auto-attribution soit effectuée avec succès :

- a) la *préservation du contenu intentionnel* de l'état occurrent d'un niveau mental à l'autre (de l'état simplement activé à la conscience réflexive de cet état);
- b) la *reconnaissance de l'attitude* ou de l'opération ayant pour objet ce contenu intentionnel : le sujet doit identifier non seulement un contenu de pensée, mais aussi une attitude à l'égard de ce contenu;
- c) la reconnaissance que ce qui est pensé l'est *par le sujet qui pense*. Il ne s'agit pas, comme l'a bien vu le théoricien de l'immunité à l'erreur d'identification, d'une erreur sur la personne; mais plutôt de l'acceptation ou du déni possibles que moi, qui m'identifie sans difficulté, aie la propriété d'être le penseur de cette pensée.

Trois arguments de niveau différent permettent de rejeter la nécessité métaphysique de la transition. Le premier consiste à s'attaquer à la relation constitutive qui est censée s'appliquer à un contenu mental et à la saisie réflexive de ce contenu. Le cas de l'expérience d'un animal non humain, qui est capable de se représenter spatialement les objets de son environnement, mais n'a pas pour autant les concepts de second ordre l'autorisant à former le contenu réflexif correspondant, paraît s'opposer à la thèse constitutive. Quoique l'on puisse supposer qu'il existe des états mentaux occurrents chez cet animal, on a des raisons indépendantes de penser qu'il ne dispose pas d'états mentaux réflexifs ni d'ailleurs du concept de concept ni d'aucun concept psychologique.

La seconde objection consiste à récuser la nécessité métaphysique de l'ampliation. Il n'est pas suffisant de penser que *P* pour dériver [*Je Y* que *P*]. Ce qui peut au plus être dérivé est [*P*, de type *Y*]. La validité de la dérivation présupposerait que le "je" soit également partie constitutive de l'état mental occurrent, ce qui n'est pas le cas. Cette seconde objection résiste ainsi à la *petitio* cartésienne par la *petitio* contraire. On peut évidemment estimer l'argument insuffisant. Il a toutefois le mérite d'indiquer une lacune argumentative dans l'argument constitutif.

La troisième objection consiste à invoquer un argument psychopathologique, que l'on peut aussi considérer comme une expérience de pensée, montrant qu'il n'y a pas de nécessité métaphysique à reconnaître que l'on a soi-même formé une pensée de premier ordre. Pour les besoins de l'argument, on suppose qu'il existe une personne qui ait l'occurrence mentale (3), et l'on montre que cette personne n'est pas métaphysiquement contrainte de s'attribuer cette pensée lors même qu'elle se pose la question de savoir qui est l'auteur de cette pensée. Rien ne s'oppose en effet à ce que la personne qui pense [*P*] --en supposant, encore une fois, qu'elle existe indépendamment de ses états de pensée occurrents -- attribue la

pensée (3) à autrui. On peut bien entendu estimer que le patient se trompe. Mais on ne peut sous-estimer l'existence et la force de cette conviction, dont la clinique de la schizophrénie offre maint exemple. Le patient souffrant d'un délire d'intrusion de la pensée nie souvent qu'une pensée occurrente soit sa propre pensée. Quoiqu'il ne se trompe vraisemblablement ni dans le contenu, ni dans le type de pensée dont il s'agit, (et manifeste de ce point de vue l'autorité de la première personne sur ses propres contenus mentaux), il conteste que cette pensée occurrente soit la sienne. Il attribue généralement la pensée intruse à l'influence d'un autre penseur qui a mystérieusement gagné le contrôle de son cerveau). Il y a dans ce cas une dissociation très claire entre la pensée en *Je* de second ordre et l'attribution à ce *Je* de la pensée de premier ordre [P]. Ce qui nous intéresse ici, c'est que cette dissociation est facilement compréhensible, et ne menace en rien la compréhension que nous avons de ce qu'est un état mental.

Du point de vue de la nécessité métaphysique, c'est ainsi essentiellement la troisième dimension de l'auto-attribution qui fait problème. Même si l'on admet que les fonctions mentales *occuper l'état occurrent* [P] / *occuper l'état occurrent* [Y que P]) sont en relation de constitution partielle, rien de tel ne vaut de l'énoncé d'auto-attribution (4), [Je Y que P]. Le référent du constituant "je" n'est pas inclus dans la pensée de l'expérience égocentrique que fait le sujet. L'ampliation ontologique qui intervient dans (4) exclut que la transition soit ancrée métaphysiquement dans (3), même si la réciproque peut être concédée.

L'hypothèse de la nécessité épistémique

Dans la longue analyse qu'il consacre au problème de la transition, Peacocke renonce lui aussi à l'hypothèse d'une nécessité métaphysique, et suggère que ce qui peut fonder la nécessité du passage de (3) à (4), étant donné que l'attitude propositionnelle est de type Y, tient plutôt à une nécessité *épistémique*. Dans toute expérience perceptive ou non, le penseur est relié à l'objet de manière à acquérir une information non-descriptive. Ainsi toute expérience, qu'elle soit visuelle, auditive, olfactive, nociceptive ou émotionnelle, implique l'usage de démonstratifs. L'idée est que les démonstratifs qui présentent l'expérience dans (1), *ce* téléphone, *cette* table, impliquent l'idée de présentation visuelle. L'usage des démonstratifs permet ainsi de restreindre les contextes de manière à rendre l'implication de (1) à (2) *valide*. Tous les contextes où (1) est vrai rendent également (2) vrai.

Pour comprendre l'argument de Peacocke, il faut rappeler que Kaplan a introduit à propos des pensées démonstratives une distinction importante entre le contenu d'une pensée (ce qu'il appelle "objet de la pensée") et la signification cognitive de cet objet, c'est-à-dire son "caractère", ou encore le mode de présentation correspondant. Ainsi, le contenu mental (1) ne contient-il pas à proprement parler le mode de présentation [démonstratif visuel] ; mais il ne prend sa *signification cognitive* pour le penseur que parce que les modes de présentation des individus "téléphone" et "table" ainsi que le contexte où la présentation intervient sont donnés. L'attitude propositionnelle de façon générale est déterminée non par le seul contenu, mais par son mode de présentation, c'est-à-dire par le caractère sous lequel la proposition est appréhendée. Quoique les modes de présentation ne fassent pas partie du contenu intentionnel, ils font partie du caractère associé à ce contenu pour former la pensée complète, celle qui fait l'objet d'une attitude propositionnelle particulière.

Comme y insiste Peacocke, cette analyse a l'intérêt de mettre en lumière le rôle de précondition que joue dans l'auto-attribution la maîtrise des mêmes concepts pour former et pour s'auto-attribuer une pensée sur le monde, c'est-à-dire la condition (a) évoquée plus haut. Par cette analyse, est en effet garantie la transition de la vérité de l'auto-attribution dans tous les contextes où l'énoncé de premier ordre correspondant est vrai, *une fois présumée l'existence d'un sujet qui les énonce ou qui les pense*. Toutefois cette analyse ne nous donne pas d'explication de la transition sur les points b et c. En ce qui concerne le point b, on ne voit pas sur quoi le sujet se fonde pour identifier l'attitude propositionnelle qui est la sienne : comment peut-il subsumer sa présente expérience sous le concept de "vision" ou de "croyance" ? Ce que l'analyse nous dit, c'est que pour que l'auto-attribution soit correcte, il faut qu'elle

soit rationnellement justifiée en partie par le fait que le sujet a bien cette expérience ou cette attitude. Mais sur quoi le sujet peut-il s'appuyer pour parvenir à cette auto-attribution ?

4 - La transition comme processus

Une première solution consisterait à dire que le sujet doit apprendre à établir qu'une instance de propriété subjective appartient à une certaine catégorie : perception, croyance, désir, etc. Il peut établir, par exemple, qu'il s'agit d'une expérience visuelle parce que les qualia qui sont associés à la saisie de la relation spatiale [être sur] sont des couleurs, des formes, etc. Comme les expériences visuelles ont en commun de provoquer des impressions subjectives distinctives, le sujet peut rassembler ces impressions dans une représentation catégorielle unique. Si, comme le soutient Goldman, il y a une impression distinctive pour toutes les attitudes, ainsi que pour le degré auquel elles s'appliquent à leur contenu, la transition entre (3) et (4) consiste dans la détection par le sujet qu'il se trouve dans l'attitude Y. Selon cette théorie, l'introspection n'est pas à proprement parler la perception d'états internes. Introspecter consiste plutôt à effectuer la mise en correspondance entre une représentation catégorielle mémorisée et une représentation instanciée ou (dans le cas des sensations) un état instancié.

Cette analyse peut être appliquée soit aux seuls contenus d'expérience, qui seraient alors individués de manière étroite par les états psychologiques du sujet, soit aux seules attitudes propositionnelles, soit aux deux. En choisissant l'une ou l'autre de ces alternatives - Goldman pour lui-même choisit la dernière - on prête le flanc aux critiques externalistes bien connues. S'il est vrai que la pensée consciente conduit rationnellement à la formation de croyances sources de connaissances, il paraît nécessaire que soient individués de manière externe le contenu des croyances *et* le processus de pensée rationnelle. Si ces deux conditions n'étaient pas remplies, on n'aurait pas la garantie que des connaissances soient produites par le processus de formation de croyance. En effet, le penseur pourrait former ces croyances sans être dans les relations adéquates avec son environnement pour que les contenus de sa conscience produisent ses croyances de manière rationnelle.

Une seconde stratégie consisterait à dire que, chaque fois qu'un penseur a une expérience particulière, il a une disposition à former automatiquement une pensée portant sur *l'occurrence* de cette expérience. Cette disposition n'est pas accessible à la personne, et ne s'appuie donc sur aucune impression distinctive. Elle survient sur un mécanisme subpersonnel, et ainsi se passe de toute raison donnée par le sujet pour former le jugement sur l'attitude qui est la sienne. Cette théorie due entre autres à Shoemaker est pour cela nommée "théorie de l'absence de raison". Comme le note ailleurs Sydney Shoemaker, la disposition ne s'exprime que lorsque certaines conditions sont données, en particulier lorsque le sujet dispose des concepts de soi-même et d'état mental, et se pose la question de savoir dans quel état il se trouve. Le sujet n'aura ainsi de connaissance de soi que pour autant que le mécanisme subpersonnel d'auto-attribution sera fiable. Distinguons deux manières de comprendre cette fiabilité. La première est essentiellement étroitement fonctionnelle. Le mécanisme subpersonnel envisagé est constitué par un dispositif neuronal assurant que, étant donné un certain type d'entrée, un certain type d'effet auto-identificateur ou auto-attributeur s'ensuivra; il ne produira l'effet attendu que pour autant qu'il fonctionne normalement, ce qui dépend d'un ensemble de conditions ayant trait à la chimie cérébrale, à l'état occurrent des neurones, des synapses, etc. Dire qu'un tel mécanisme est fiable, c'est dire qu'il produira subpersonnellement une information qui pourra finalement être utilisée par le sujet sans qu'il connaisse le mécanisme auquel il doit son savoir.

Toutefois, la fiabilité peut être étendue au-delà des caractéristiques fonctionnelles, causales et donc physiques du mécanisme considéré si le sujet lui-même *est capable d'apprécier la valeur informationnelle de la sortie*. Utilisant son intelligence et ses capacités conceptuelles, le sujet peut le cas échéant rejeter l'impression formée subpersonnellement. En voici un exemple. Il n'est pas rare que les montagnards non-entraînés à la raréfaction de l'oxygène en très haute altitude soient sujets à des hallucinations. L'un d'entre eux rapporte avoir eu, à 5000 mètres, la "vision" de petits hommes marchant au pas. "J'ai l'impression de percevoir que P, se dit alors le sujet, mais il est impossible que je voie [P] parce que [P] est sinon impossible, du moins très improbable. Je dois halluciner". La fiabilité causale du mécanisme a dans ce cas été complétée par la fiabilité de l'ensemble du système rationnel de

l'auto-attribution.

Du point de vue qui nous intéresse, à savoir l'existence d'un lien épistémique garantissant que la transition de (3) à (4) est source de connaissance, il faut bien constater que l'invocation d'un mécanisme subpersonnel peut certes guider de manière fiable l'identification du type de l'attitude Y que le sujet a la disposition de s'attribuer lorsqu'il forme la pensée de contenu [P], et ainsi garantir la connaissance de l'attitude en question. Mais évidemment, ce qui n'est nullement garanti dans cette approche, c'est que le sujet attribue cette attitude ou cette opération mentale (croire, percevoir, rêver, halluciner, etc) *au sujet même qu'il est*. Encore une fois, il ne s'agit pas de plaider pour l'inexistence du sujet, mais de maintenir que rien, dans les analyses envisagées, n'interdit la possibilité qu'un sujet qui s'identifie correctement comme lui-même puisse nier être le penseur de cette pensée.

Toutes les analyses épistémiques envisagées échouent à satisfaire la condition c évoquée plus haut : elle permettent au mieux de justifier l'implication épistémiquement nécessaire non de (3) à (4) mais de (3) à (5), c'est-à-dire à :

(5) Maintenant, est formée une attitude propositionnelle ayant le contenu P sous une perspective mienne.

En toute rigueur, le caractère implique le sujet qui pense cette pensée uniquement de manière épistémique, c'est-à-dire à travers la pertinence égocentrique du contenu pensé. Le caractère n'est qu'un mode de donation de la pensée, et non un mode d'introduction d'un individu qui penserait la pensée : que le caractère soit subjectif n'implique en rien le titre qu'aurait un sujet à s'attribuer l'expérience considérée.

Une autre manière de présenter cet argument consisterait à invoquer l'existence de nombreuses illusions concernant la conscience du mien. Quoique les philosophes qui s'appuient sur la conscience pour dériver l'auto-attribution défendent généralement une conception externaliste des contenus mentaux, on peut objecter que précisément la conscience d'être en relation subjective avec quelque chose (qui s'exprime dans l'usage adjectival traduit plus haut par "sous une perspective mienne") n'a pas toujours d'ancrage dans deux termes indépendants objectifs, dont l'un serait le sujet, et l'autre le contenu mental. Par exemple, contemplant les occupants d'un manège dans une fête foraine, il est banal d'éprouver soi-même l'impression que l'on a lorsqu'on est en haut du "Grand Huit". Utilisant son intelligence et ses capacités conceptuelles, on peut bloquer l'auto-attribution que l'on perçoit soi-même le vide.

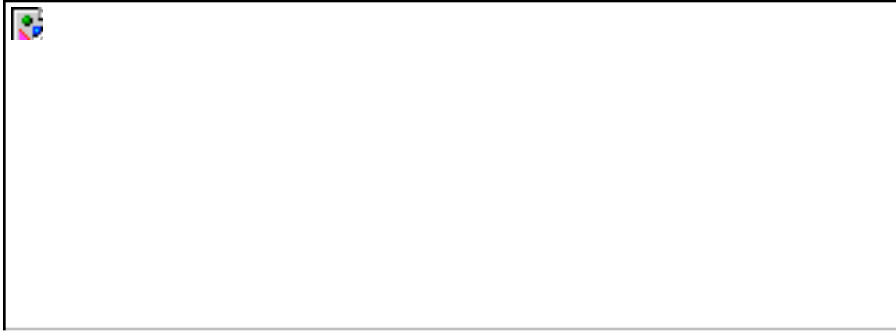
S'il n'est pas possible de tirer du caractère davantage qu'un rapport adverbial avec le sujet putatif de l'expérience, rien ne garantit que ce rapport adverbial ne soit rejeté par le sujet de la pensée, utilisant ses capacités propres d'intelligence et de conceptualisation.

5 - L'hypothèse du lien rationnel

Selon Christopher Peacocke, le contenu de l'état conscient de premier ordre ne fournit pas une *donnée* permettant d'appuyer la transition ; mais il fournit néanmoins une *raison* d'effectuer cette transition ; l'auto-attribution a la propriété d'être une *transition primitive rationnellement imposée* au penseur qui a l'état conscient de premier ordre du fait qu'il dispose des concepts pertinents. Il se range ainsi parmi les théoriciens qui défendent l'existence d'un lien épistémique nécessaire entre (3) et (4). Nous avons présenté plus haut cette position, et avons vu que, lorsque nous développons les points laissés en suspens quant aux raisons que le sujet peut avoir de s'auto-attribuer ses états mentaux de premier ordre, nous ne parvenons pas à découvrir dans les termes de cette théorie la nécessité épistémique de la convergence référentielle entre une personne supposée donnée et le sujet éventuel possédant les états mentaux occurrents.

Il peut être utile, avant de proposer une théorie de cette convergence référentielle, de rappeler dans quels termes, de l'avis de Peacocke, une théorie satisfaisante de la transition doit généralement se présenter. Il

propose non une théorie substantielle de la transition, mais un schéma d'explication, c'est-à-dire une classe de théories qu'il nomme *théories - delta*, ainsi nommées parce que, prenant pour bases l'état mental de contenu [P] et de type y et le jugement [Je y que P], elles associent à chacune de ces bases l'unique sujet auquel le terme "je" fait référence et qui possède l'état mental de contenu [P].



Ce schéma part de deux prémisses fondamentales, que l'on peut expliciter dans les propositions (6) et (7) :

(6) Pour tout état ou événement mental conscient, il existe un et un seul possesseur de cet état ou événement, qui est le sujet conscient.

(7) L'occurrence de l'auto-attribution d'un état mental conscient est en relation de co-conscience avec l'occurrence de l'état mental conscient correspondant.

Etant donné (6) et (7), le delta paraît pouvoir être fermé en affirmant l'identité *a priori* entre le possesseur de l'état conscient occurrent de contenu p et de type y et la référence de *Je* dans "Je y que p ". Etant donné (7) qui affirme le caractère co-conscient de l'état et de son auto-attribution, et par application de l'unicité du porteur affirmée dans (6), on peut apparemment déduire que l'auto-attribution, étant également l'objet d'une occurrence d'état conscient - est possédée par le sujet. Comme on va le voir, il reste à établir toutefois que ce porteur commun des *deux* états co-conscients est également l'objet auquel fait référence le *Je* dans la pensée auto-attributive.

La pointe du delta donne évidemment pour ainsi dire l'emplacement de la solution recherchée : le sujet qui possède ces états est, en un sens à préciser, *a priori* identique au sujet auquel fait référence la pensée auto-attributive [Je y que p]. Comment la convergence entre la référence à la première personne et le propriétaire de l'état mental est-elle assurée ? En d'autres termes, comment garantir la relation de co-conscience sans présupposer déjà l'existence d'un seul porteur pour deux états conscients différents ? Peacocke invoque la démonstration fré géenne selon laquelle un état conscient implique nécessairement l'existence d'un *porteur* de cet état. Toutefois même si nous avons la certitude qu'il existe un porteur pour tout état conscient, il n'est pas par là-même garanti que ce porteur soit identique au sujet qui s'auto-attribue le contenu de conscience. Comme l'écrit Shoemaker dans un contexte voisin, "Pour pouvoir m'identifier moi-même comme étant moi-même par la possession de *cette* propriété [ici : la propriété d'être le porteur de l'état conscient p de type y], il faudrait que je sache que *je* l'observe par le sens interne, et *cette* connaissance de soi, étant le fondement de mon identification de moi comme moi-même, ne pourrait pas être elle-même fondée sur cette identification". Peacocke est parfaitement conscient de cette difficulté, et rejette à juste titre toute conception de la transition qui la fonderait sur une observation quelconque.

Affirmer que seule la présupposition de l'existence d'un sujet auquel il est fait référence par les pensées en *je*, sujet qui serait identique au possesseur des états mentaux conscients, permet de fonder rationnellement les auto-attributions et par là, la connaissance de soi, peut constituer un argument transcendantal si l'on dispose déjà de la preuve indépendante de l'objectivité de la connaissance de soi. Mais il n'a pas la valeur d'une preuve directe de l'unité du possesseur d'états conscients et du *je* qui

s'attribue ces états.

Nous pouvons tirer les enseignements des discussions précédentes en rappelant deux thèses classiques. La première rappelle les conditions quasi-indexicales requises par l'auto-attribution. Pour atteindre la connaissance de soi-même comme étant soi, il faut davantage qu'une simple identité entre un je et l'attribution d'une propriété particulière à ce je. Comme Hector-Neri Castaneda l'a montré, "les propositions sur un *Je* donné ne peuvent être des objets pleins de croyance que si la croyance en question appartient au même *Je*". L'usage normal du pronom personnel *je* ne garantit pas cette commune appartenance. L'usage quasi-indexical, qu'il note *je** (*il**, *nous**, *etc.*), marque l'identité dans le discours oblique de l'auto-attribution entre le sujet qui forme la croyance et l'objet -le sujet lui-même - sur lequel porte cette croyance. Sans cette capacité de faire référence quasi-indexicalement au moyen d'une désignation qui engage réflexivement l'usage précédent du pronom personnel, on pourrait certes acquérir des éléments d'information portant en fait sur soi, mais non avoir la connaissance réflexive correspondante. C'est à utiliser le quasi-indexical dans le cas de la proposition (4) que nous nous efforçons jusqu'à présent sans succès.

Le second enseignement, c'est que même à supposer que l'existence d'états conscients provoque causalement ou conditionne épistémiquement la transition vers un état d'ordre supérieur, un nouvel argument doit être produit pour tirer de ce qui ne peut être qu'une notation adverbiale de cette pensée de second ordre - comme dans (5) - une relation d'appartenance à un sujet.

Dans ce qui suit, nous tenterons de montrer que la seule façon de justifier l'unicité du sujet de la pensée comme porteur de ses états mentaux consiste à examiner dans quelles conditions effectives ce sujet pense, et est amené à former des pensées sur le fait qu'il pense. L'examen de ces conditions fait ressortir que le sujet pense pour agir. La thèse que nous défendrons consiste à tirer la métaphysique du sujet de la capacité même de former des états mentaux de second ordre pour réguler son action.

6 - L'appropriation des états mentaux et l'agir mental

C'est la transition qui m'attribue à *moi-même* - au sens quasi-indexical de Castaneda - le contenu mental occurrent qui nous intéresse maintenant, la transition proprement appropriative (par opposition à la transition du premier au second ordre). Non plus le fondement du passage d'un état mental conscient à la pensée de cet état mental (et non d'un autre) , mais à la pensée que c'est moi qui l'ai.

Les difficultés précédentes peuvent être diagnostiquées comme provenant de la manière exclusive dont le *je* a été analysé : en relation avec le seul contenu informationnel de ses propres pensées. Il ne suffit pas d'examiner les conditions où le sujet s'auto-attribue un contenu mental donné pour obtenir le fondement de son unicité. L'unicité dont nous parlons ne peut-être seulement momentanée, liée à l'existence d'un instant unique pendant lequel divers contenus de pensée sont co-conscients. Car rien ne peut apparemment interdire que se constituent en parallèle autant de sujets que d'états conscients ou de paires de tels états, étant donné la dépendance entre "être le possesseur d'un état conscient" et la relation de co-conscience entre les états mentaux possédés par un seul possesseur. Rien non plus ne vient fixer les conditions temporelles d'individuation des personnes. Combien d'instant successifs une personne doit-elle occuper pour être une personne ? L'une des erreurs que nous avons commises réside dans l'idée que c'est dans la seule conscience occurrente de ses propres états mentaux que réside le principe de l'unicité du possesseur de la pensée.

Nous avons des raisons indépendantes d'élargir les bases sur lesquelles examiner la question de l'unicité du "je", et de prendre en compte, outre ce que le sujet sait originellement de lui-même, des connaissances rapportées au sujet -- dont il n'est pas lui-même la source et qui sont formulées en troisième personne, ainsi que les autres applications du concept de sujet. Ces raisons tiennent à l'exigence générale que doit remplir tout concept, à savoir la *contrainte de généralité*. Pour comprendre le concept de "je", il faut pouvoir faire varier les deux "séries de pensées" qui entrent dans la

compréhension de "J'ai *F*". La première consiste dans la série "j'ai *G*", "*x* a *G*", "*y* a *G*". Comme l'écrit Strawson, "Une condition nécessaire pour que quelqu'un s'attribue à lui-même des états de conscience et des expériences, comme il le fait, est qu'il puisse également les attribuer, ou qu'il soit disposé à les attribuer, à d'autres que lui-même". Gareth Evans ajoute à cette première condition la disposition à comprendre une deuxième série : "j' ai *F*", "j' ai *G*", "j'ai *H*", etc. Comme l'observe Evans, "Nous sommes parfaitement capables de saisir des propositions nous concernant que nous sommes entièrement incapables de dire vraies ou fausses, ni même de commencer à justifier. Je peux comprendre la pensée que j'ai été nourri au sein, par exemple, ou que j'ai été malheureux le jour de mon premier anniversaire". L'idée qu'Evans élabore ici, c'est que pour pouvoir avoir l'idée de soi comme sujet qui s'auto-attribue une connaissance acquise sur la base de son expérience consciente, il faut pouvoir saisir l'identité entre le "je" ainsi visé et d'autres identifications par autrui de la personne à laquelle "je" fait référence, sur la base de propriétés que la personne en question n'a pas connues directement.

Les propriétés pertinentes pour l'unicité du *je* de l'auto-attribution seront alors tantôt psychologiques, tantôt physiques ; tantôt auto-attribuées, tantôt rapportées par autrui. Le fait que ces propriétés ne soient pas toutes construites sous le point de vue du sujet garantit que je puisse me considérer comme un objet du monde objectif, une personne parmi d'autres. Non pas au sens où les autres me voient et me constituent comme personne - car je peux être une personne même sans que cette reconnaissance me soit *de facto* accordée. Mais au sens où les prédicats qui sont auto-appropriés par un sujet sont des propriétés quelconques qui élèvent leur possesseur à l'objectivité (l'indépendance à l'égard de ma pensée et de celle d'autrui). Les propriétés physiques, de type spatio-temporel, comme [être né à Lausanne le 1er octobre 1900] fournissent un ancrage objectif qui font de la personne un objet du monde parmi d'autres. Pas plus qu'on ne peut identifier d'autres sujets en les identifiant uniquement comme sujets d'expériences, comme possesseurs d'états de conscience, comme l'observe Strawson, on ne peut s'identifier soi-même comme unique sur la base de ses seules propriétés mentales. Pour résumer la stratégie proposée par Evans, la meilleure façon de garantir l'unicité du "je" à travers la série de ses propriétés mentales auto-attribuées consiste à traiter ces auto-attributions comme un cas particulier d'attributions, et de saisir le concept de "je" dans le réseau plus général que ces attributions constituent.

Même si cette stratégie atteint le but recherché, qui est de saisir le "je" comme unique à travers les manifestations de sa conscience réflexive, il n'est pas sûr qu'elle soit la seule, ni la plus économique. Si elle est suffisante, la contrainte de généralité n'est peut-être pas nécessaire pour l'unicité du "je" qui se connaît comme possesseur d'états intentionnels. On peut en effet objecter que le sujet au sens plein d'Evans, celui qui est capable de faire le lien entre les propriétés qu'il s'attribue et celles qui lui sont attribuées, représente une forme achevée de personne qui n'est pas nécessairement mise en jeu par l'unicité du Cogito. Plus profondément, on peut douter que les propriétés mentales soient une base *insuffisante* pour fonder l'identité du "je" qui pense. Car comment un tel sens de soi pourrait-il être ancré, si ce n'est en définitive dans des croyances et des motivations ? Pour pouvoir revenir mieux armé sur cette possibilité, il faut procéder à un second recentrage de l'enquête, en explorant cette fois les contextes concrets où s'effectue mentalement la réidentification du "je".

Notre première erreur était de nous être attachés exclusivement aux états occurrents. La deuxième erreur que nous avons commise consiste dans le fait de n'avoir considéré que les seules *attitudes propositionnelles* occurrentes du sujet, dans la transition entre :

(3) Maintenant [P] (de type Y)

et

(4) Maintenant je Y que P

Une théorie alternative concernant le type de propriétés mentales pertinentes pour la réidentification d'un je identique consiste à examiner non pas une attitude occurrente, mais le réseau dont elle fait partie dans le raisonnement pratique. Esquissons les grandes lignes de cette théorie, avant de revenir plus loin sur les arguments qui plaident pour elle. L'unicité du "je" dans (4) ne se comprend que s'il existe plusieurs

attitudes propositionnelles, y, c, f, x, qui font ou ont fait l'objet d'une auto-attribution, et si ces attitudes propositionnelles ont formé un réseau de raisons ou ont été conjointement exploitées dans le cadre d'une action déterminée. La mise en perspective temporelle et fonctionnelle de cette série d'états intentionnels permet de conférer un sens plus clair à l'unicité du sujet possesseur de tous ces états. Ce qui réunit tous ces usages du "je" comme sujet d'auto-attributions réside non pas véritablement dans l'observation que les diverses attitudes propositionnelles sont *causées* par le même agent, mais plus profondément dans le fait que le même individu est *concerné* par ces attitudes propositionnelles fonctionnellement distinctes, dans la mesure où elles ont dirigé son action, et où elles ont impliqué un engagement rationnel. A la différence du porteur d'une pensée, dont l'individuation spatiale et temporelle est floue, et dont le rôle fonctionnel peut paraître superflu, c'est-à-dire de type "homonculaire", l'agent d'une action peut être individué plus clairement par les propriétés fonctionnelles de l'action.

Cette nouvelle hypothèse sur le "je" ne consiste pas à identifier simplement le "je" et l'"agent" de l'action. Elle s'appuie sur la vérité banale - mais essentielle - selon laquelle les diverses attitudes propositionnelles, perceptions, croyances, désirs, espoirs, qui sont réflexivement présentées dans les pensées de second ordre correspondantes, constituent les éléments déterminants d'une action individuelle d'un type particulier, que nous appellerons "l'action réfléchie". C'est au niveau de l'action réfléchie qu'est jugée l'intégration relative des attitudes propositionnelles entretenues par l'agent, et c'est parce que les croyances et les désirs ont pour finalité une action individuelle accomplie dans un contexte particulier donné que les attitudes propositionnelles doivent être hiérarchisées et rendues cohérentes entre elles. L'unicité de la personne constitue alors le versant normatif et stable au fil du temps de l'unicité de l'agent au fil de ses actions.

Dans la présente perspective, l'agent ne joue pas véritablement de rôle causal dans l'action : ce sont les états intentionnels de l'agent qui occupent exclusivement ce type de rôle. On ne peut pas, sur la base de cette intuition, tout simplement réduire le sujet identique à soi à un sous-ensemble de ses états mentaux - à ceux qui auraient joué le rôle causal en question. Car ce sous-ensemble ne cesse de se modifier, et l'on ne disposerait dans ce cas d'aucune capacité réidentificatoire du "je". En revanche, il est possible, dans cette analyse, de considérer que le "je" est constitué par l'engagement normatif propre à tout agent ayant une information réflexive sur ses propres états à *réviser* ses propres dispositions si une incohérence ou une inadéquation moyen-fin apparaît et, plus généralement, à répondre rationnellement de ses actions en faisant valoir le contenu de ses attitudes dans leur rapport avec l'action envisagée. Il y a ainsi un lien direct entre la capacité de former des états mentaux de second ordre, l'aptitude à réviser les modalités de son action, et l'engagement envers ses propres attitudes ainsi révisées rationnellement. Ce qui, dans cette perspective, donne les conditions nécessaires de la réidentification du sujet comme le même, réside dans la persistance d'un engagement à l'égard d'un ensemble d'attitudes propositionnelles, cet engagement se traduisant par la disposition à réviser ses attitudes le cas échéant.

Une expérience de pensée inspirée par Akeel Bilgrami permet d'établir et de préciser le rôle de l'action dans la réidentification du "je". Supposons un être nommé Oblomov, comme le célèbre personnage du roman d'Ivan Gontcharov, qui aurait la propriété d'être *entièrement* passif. Même s'il a des états mentaux conscients, supposons que rien de ces états mentaux ne passe par un contrôle, un acte d'attention, une orientation délibérée, une sélection. Supposons en outre qu'il n'ait pas la faculté de faire valoir des préférences, ni de viser des objectifs particuliers. Supposons que sa pensée soit elle aussi entièrement réactive : que les pensées se produisent dans son esprit sans qu'il ait rien fait pour les produire ou les retenir. Peut-on dire qu'un tel individu puisse atteindre une connaissance de soi ?

Il ne le peut pas pour une raison massive : il ne vaut pas la peine de fixer des croyances ou des désirs si l'on ne peut pas agir sur leur base. Ce n'est pas là un fait empirique. C'est un aspect conceptuel constitutif des états mentaux, qui nous oblige à réviser notre description initiale d'Oblomov. Si des croyances et des désirs pouvaient être fixés sans faire aucune différence pour le comportement individuel, ces croyances et désirs seraient dépourvus de toute valeur normative, et seraient par conséquent des indicateurs non fonctionnels. C'est en exerçant ses dispositions à agir que l'on sélectionne les désirs efficaces sur la base des croyances et des préférences que l'on a. Un sujet qui serait exclusivement mû par un mécanisme causal ou par un agent extérieurs, n'aurait que faire d'états intentionnels, et *a fortiori* d'un "je".

Imaginons maintenant qu'un autre sujet, Oblatov, soit affecté d'une incapacité un peu différente de celle d'Oblomov. Au lieu d'être entièrement passif, il peut agir, mais uniquement sur la base de ses états de premier ordre. Incapable de former réflexivement des pensées de second ordre, (et être informé du fait qu'il croit, qu'il perçoit, ou qu'il désire), il ne peut pas non plus former des désirs de second ordre sur ses états intentionnels de premier ordre : Désirer s'informer davantage, rétablir la cohérence dans ses préférences, faire porter ses désirs de premier ordre sur des objets dignes d'être possédés, etc.

Notons ici que ce qui fait défaut à Oblatov ce n'est pas, comme chez Oblomov, toute capacité à agir. C'est seulement la capacité à agir d'une manière réfléchie. Oblatov est un agent, et utilise donc ses croyances et ses désirs pour agir. Mais Oblatov manque d'une capacité cruciale : celle de pouvoir rendre raison de ses choix. Il ne peut pas le faire parce qu'il n'a pas accès au contenu de ses propres états : il n'a pas de croyances sur ses désirs de premier ordre. Si un tel sujet avait des croyances incohérentes et qu'il eût conscience de cette incohérence, il devrait penser que "ce sont les *faits* qui sont incohérents", comme le note Shoemaker dans un autre contexte.

Il est intéressant de noter qu'Oblatov n'est pas pour autant entièrement dépourvu de toute capacité de délibérer, si l'on entend par là la capacité de faire émerger le plus puissant de ses désirs. Le mode de délibération qui lui est fermé consiste seulement dans la délibération conçue comme "entreprise autocritique", impliquant la reconnaissance de l'éventualité que certaines croyances soient fausses, et de la nécessité de procéder à de nouveaux tests.

La réflexion sur ce cas permet de dissocier clairement la capacité d'agir de l'existence d'un "je". Même si l'on admet qu' Oblatov est un agent, il n'est évidemment pas une personne, un sujet réidentifiable, parce qu'il est constitutif d'une personne d'avoir la disposition à rendre raison de ses actions. Oblatov ne peut pas en rendre raison parce qu'il n'a pas accès à ses propres états mentaux. Remarquons que, dans cette hypothèse, la personne n'est pas une *condition* de l'action, -- les attitudes propositionnelles appropriées le sont, ainsi que l'intégrité du système perceptivo-moteur -- mais une *conséquence* de la capacité, développée à *la faveur* de l'action, de rendre raison de son action sur la base du contenu de ses attitudes propositionnelles. L'unicité de la personne est le produit du fait que l'action se trouve *de facto* ne concerner qu'un seul agent, celui qui a mis en jeu ses propres attitudes propositionnelles pour réguler son propre comportement. Il s'ensuit une conséquence que l'on peut juger étrange. Si, dans une société anti-individualiste, plusieurs agents répondaient régulièrement de leurs actions -- exclusivement collectives -- en invoquant des attitudes propositionnelles distribuées entre eux, il faudrait dire qu'il s'agit d'une seule personne. Ce qui, dans cette conséquence, est difficile à admettre, consiste non dans l'identification d'une "personne morale" - conséquence théoriquement intéressante - mais dans l'absence de rapport entre raisons données et comportement subséquent. Les conditions de la réidentification de la personne sont alors entièrement tournées vers le passé, et le rapport entre action et normativité est alors perdu.

Une troisième direction d'enquête devrait nous permettre de résoudre cette difficulté. Il s'agit cette fois d'exploiter le fait que les attitudes propositionnelles, lorsqu'elles sont révisées réflexivement, doivent faire l'objet d'un contrôle, suivi des transformations rationnelles *correspondantes* (visant à supprimer les contradictions, réviser les préférences, etc.). Or ce contrôle qui permet à l'individu de réajuster ses états mentaux aux normes de la rationalité, correctement compris, doit nous offrir une prise nouvelle sur la question de la réidentification de soi. L'idée nouvelle sur laquelle s'appuie cette précision de l'hypothèse, consiste à dire qu'un sujet se réidentifie à travers *les actions mentales* qu'il effectue, où il est à la fois *évaluateur* de ses attitudes propositionnelles, *agent* des modifications qu'il aura globalement estimées nécessaires, et *support* des propriétés auto-affectées. Nous avons maintenant une triple condition qui garantit la possibilité d'une réidentification par la reconnaissance de la coïncidence (*dans la même action mentale*) entre l'agir et le pâtir, c'est-à-dire d'une auto-transformation. Expliquons nous.

Une action mentale peut être définie comme un type d'action qui, comme toute action, est causé par des désirs et des croyances, et dont le contenu intentionnel est d'obtenir une propriété mentale nouvelle en utilisant à cet effet des moyens qui sont eux-mêmes mentaux. Ainsi défini, ce concept d'action mentale ne s'applique généralement pas aux opérations de premier ordre, telles que percevoir *P*, croire *Q*, désirer

R. Car ces opérations de premier ordre n'ont généralement fait l'objet d'aucun contrôle; elles n'ont pas été effectuées en vue d'obtenir une certaine propriété mentale. Elles se sont produites dans l'esprit du sujet sous l'influence de causes variées, endogènes ou exogènes. En revanche, l'attention contrôlée constitue une action puisqu'elle se développe sur la base de la croyance qu'il existe une propriété mentale qui mérite d'être atteinte (une nouvelle représentation perçue, que seule l'attention permet de former). Percevoir attentivement que *P* se distingue ainsi fondamentalement de percevoir que *P*. De même, le souvenir contrôlé, par opposition au souvenir automatique, constitue un acte mental. Chercher à se souvenir d'une date se distingue du rappel automatique d'une date. Modifier délibérément ses désirs ou ses préférences à la lumière de ses connaissances et de ses valeurs globales est une variété d'action mentale, en vue de rendre efficaces de nouveaux désirs, ou de nouvelles échelles de préférences. Comme y a insisté Harry Frankfurt, une différence capitale pour le concept de responsabilité oppose les désirs de premier et de second ordre : les désirs de second ordre sont indispensables pour contrôler l'opération des désirs de premier ordre.

Revenons au problème de la réidentification munis du concept d'action mentale. Il semble possible d'assembler enfin le porteur d'un état conscient et le sujet qui s'auto-attribue cet état conscient en invoquant le fait que le premier est la cible du second, tandis que le premier est le témoin et le guide de cette transformation: c'est le porteur d'un état conscient que l'action mentale cherche à modifier; ce porteur sera directement *affecté* par la modification : s'il s'agit d'attention, il percevra un état de choses qu'il n'aurait pu percevoir sans cette action. S'il s'agit de désir, il désirera ce qu'il ne désire pas encore (ou cessera de désirer ce qu'il désire). En outre, l'action mentale s'appuie comme toute action sur un feedback, sur un retour informationnel qui permet d'évaluer si l'action a été ou non réussie : il est incohérent de continuer à faire attention si la propriété perceptive a été saisie, ou de désirer si la propriété motivationnelle a été atteinte. L'auto-attribution ne peut donc s'arrêter à une simple et unique attribution. Elle doit resaisir l'auto-attribution dans le cadre de l'action mentale dont elle constitue un moyen. La réflexivité des états mentaux reflète la propriété de l'action de prélever l'assurance que les conditions de satisfaction de l'action mentale sont réunies, ou en voie de l'être. La référence au sujet dans la transition entre une simple opération mentale et son auto-attribution est indissociablement une référence au sujet qui est le porteur d'un état conscient, qui évalue ses états et les modifie parce qu'il est concerné par cette modification.

Si cette analyse est correcte, nous pouvons la vérifier rapidement sur le cas de notre expérience de pensée. Supposons maintenant un personnage, Oberov, capable d'agir physiquement, capable de former des états de second ordre de manière automatique, mais incapable d'agir mentalement : cet individu ne peut pas viser l'obtention d'une propriété mentale quelconque : il ne peut pas faire attention, souhaiter connaître, chercher à se rappeler, espérer voir, craindre de s'ennuyer, etc. A la différence d'Oblomov, et à l'instar d'Oblatov, Oberov peut agir. A la différence d'Oblatov, des pensées de second ordre peuvent se former en lui. Il ne peut pas plus tirer parti de ces pensées de second ordre qu'Oblomov de ses pensées de premier ordre, mais on doit lui reconnaître une intentionnalité du fait qu'il utilise ses pensées de premier ordre pour guider son action. Oberov peut-il dire "je" en faisant référence par là au porteur des états conscients qui sont les siens quand il a des pensées de second ordre ? Dans la mesure où Oberov manque des moyens cognitifs de s'auto-affecter mentalement, rien ne garantit l'unification de ses diverses auto-attributions. Oberov, proche des primates non-humains, ne peut pas former de concepts psychologiques. Il ne peut pas reconnaître la causalité du mental, ni l'efficacité des croyances dans le comportement. Son concept de "je" sera dépourvu de la normativité inhérente à la capacité de réviser ses jugements : il ne fera pas référence à une personne stable à travers les contextes de l'action.

On pourrait ici avancer que la solution proposée présente un inconvénient majeur, dans la mesure où elle interpose entre le penseur et le contenu de ses états mentaux de premier ordre, individués de manière externe, un niveau opaque constitué par les états mentaux occurrents de second ordre et leur gestion active. Ne perd-on pas dès lors le principe même de la rationalité de l'acquisition des croyances ? Il y a plusieurs façons de répondre à cette objection. La première consiste à observer que l'identification du sujet est *postérieure* à la dérivation d'un état de second ordre relativement à l'état de premier ordre correspondant. Ainsi, même si la construction du sujet faisait intervenir des considérations de type internaliste, elle ne compromettrait pas le caractère externaliste des contenus mentaux de premier et de

second ordre. Toutefois, on peut objecter que cette réponse n'est pas convaincante, car même si l'identification du sujet est logiquement dépendante de la capacité à effectuer la transition entre des états mentaux de premier et de second ordre, cette identification n'en commande pas moins la portée de la connaissance de soi. Faudrait-il admettre que la connaissance de soi soit d'une variété fondamentalement différente de la connaissance du monde extérieur ?

La seconde manière, beaucoup plus prometteuse, consiste à soutenir que le "je", justement parce qu'il est l'entité concernée par le succès des actions mentales, et l'enjeu de la stratégie dans laquelle les attitudes propositionnelles sont inscrites, doit être individué en partie de manière externaliste. Ce qui, dans la théorie proposée, fait d'un "je" un "je", n'est pas qu'il se reconnaisse comme tel, ou qu'il ait accès à des qualia particuliers ni même à des faits concernant la vie mentale. Ce statut dépend de l'imbrication fonctionnelle entre des attitudes de second ordre qui sont toujours elles-mêmes individuées de manière externaliste. Par exemple, dans le souvenir dirigé, l'agent constate qu'il ne peut pas spontanément se souvenir de *P*. Il se demande alors s'il *peut* s'en souvenir de manière contrôlée, et il s'engage dans le processus de remémoration de *P* en traitant toutes les inférences dont il dispose qui impliquent *P*. Rien dans ce processus ne dépend crucialement de l'impression de savoir, considérée indépendamment de la question de l'impression de savoir *que P*. Les qualia métacognitifs, si l'on peut parler ainsi des phénomènes comme "avoir quelque chose sur le bout de la langue", ou l'impression de familiarité, ne se différencient pas des qualia cognitifs : ils renvoient à des propriétés du monde (en l'occurrence, des processus cérébraux). L'information cruciale qu'apporte la métacognition concerne précisément le degré de fiabilité du système cognitif ou motivationnel relativement à un objectif de connaissance ou d'action. On ne peut rejeter cette source métacognitive d'information sur la fiabilité et la disponibilité des processus de traitement sans mettre en péril tout le processus de construction rationnelle de nos connaissances du monde extérieur, dont d'ailleurs le cerveau fait lui-même partie.

CREA, Ecole Polytechnique

1, rue Descartes

75005 Paris

proust@poly.polytechnique.fr